Multi-Disciplinary Publishing Institute Pakistan **Vol.** 4, **N0.** 1

Submitted: 12 MAY 2025 **Accepted: 8 JUL 2025** **Published: 9 JUL 2025**

Using AI to Integrate Multiple Omics to Predict the Mechanisms of Rare Diseases

Review Article

Mukarram Sharif [1*], Jack Niedzialek[2], Iqra Khan[3], Zuhera Khan[4]

[1]Department of Microbiology and Molecular Genetics, University of Okara, Renala Khurd, Pakistan.

[2]Student of STEM Academy, Clifton High School, 333 Colfax Ave, Clifton, New Jersey, USA.

[3]Assistant Professor, FCPS, Plastic Surgery, Jinnah Sindh Medical University, Pakistan.

[4]FCPS, Plastic Surgery Consultant, Patel Hospital Karachi, Pakistan.

Corresponding author:
m.mukarramsh@gmail.com

Citation

Sharif, M., Niedzialek, J., Khan, I., & Khan, Z. (2025). Using AI to integrate multiple omics to predict the mechanisms of rare diseases. *Open Access Public Health and Health Administration Review*, 4(1), 9-15.

WEBSITE: www.mdpip.com
ISSN: Print: 2959-619X
ISSN: Online: 2959-6203
PUBLISHER: MDPIP

Abstract

Rare diseases, which affect less than 1 in 2,000 people, are difficult to diagnose and treat due to their complex pathophysiology and genetic variation. Recent advances in artificial intelligence (AI) and multi-omics technologies, including proteomics, metabolomics, and genomics, may help us better understand the molecular mechanisms underlying these diseases. This study was aimed to explore how AI integrates multi-omics data to identify biomarkers, forecast treatment targets for rare diseases, and uncover disease-causing pathways. It highlights how AI may help with data complexity, enable personalized care, and enhance predictive modeling. Despite progress, problems like consistency, model interpretability, and data scarcity persist. By integrating recent research, this paper proposes future directions to accelerate clinical translation and highlights AI-driven multi-omics as a groundbreaking approach to understanding the mechanisms underlying uncommon diseases.

Keywords: AI, Multiple Omics, Rare Disease, Mechanism, Diagnose.

Introduction

Rare diseases, which affect millions of people globally, are characterized by a variety of clinical symptoms, genetic mutations, and a lack of effective treatments (Nguengang Wakap *et al*., 2020). Due to an incomplete understanding of their molecular mechanisms, most of the more than 7,000 rare diseases that have been identified do not have effective treatments (Haendel *et al*., 2020). Multi-omics methods that integrate genomics (DNA sequences, mutations), proteomics (protein expression, modifications), and metabolomics (metabolite profiles) offer a thorough understanding of biological systems by exposing interactions across molecular layers(Subramanian, Verma, Kumar, Jere, & Anamika, 2020). However, the high volume, heterogeneity, and dimensionality of multi-omics data pose significant analytical challenges that necessitate the use of advanced computational tools (Kang, Ko, & Mersha, 2022). Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), enables the integration of multi-omics data to identify disease mechanisms, biomarkers, and treatment targets (Reel, Reel, Pearson, Trucco, & Jefferson, 2021). Neural networks and graph-based approaches are examples of artificial intelligence (AI) models that can uncover non-linear patterns and molecular interactions that traditional statistical methods usually miss (Wekesa & Kimwele, 2023). Transfer learning and federated learning are AI-driven methods that leverage knowledge from larger datasets or related diseases to address data scarcity in rare diseases, where patient cohorts are small and data is sparse (van Tuijl *et al*., 2024). Recent studies have integrated multi-omics data for diseases such as spinal muscular atrophy (SMA), cystic fibrosis, and rare cancers using artificial intelligence (AI) to identify new pathways and possible treatment targets. Despite these advancements, there is still a substantial gap in the literature because no comprehensive review has summed up AI's role in rare disease multi-omics integration (Yang *et al*., 2024).

To predict the mechanisms underlying rare diseases, this review looks at how AI helps integrate genomics, proteomics, and metabolomics. It looks at AI approaches, how they are used in rare disease research, and issues like clinical translation, interpretability, and data integration. The article seeks to close this gap to direct future studies and the creation of treatments for uncommon illnesses.

Literature Review

Multi-Omics in Rare Disease Research

Over 80% of rare diseases have been linked to mutations found through genomic sequencing, making them primarily genetic (Boycott *et al*., 2017). Rare variations, such as single nucleotide polymorphisms (SNPs) or copy number variations (CNVs), linked to conditions like Duchenne muscular dystrophy (DMD) or Rett syndrome are found through genomics using whole-exome sequencing (WES) and whole-genome sequencing (WGS) (Bamshad *et al*., 2011). By analyzing protein expression and post-translational modifications, proteomics enhances genomics by exposing the functional effects of genetic mutations (Aebersold & Mann, 2003). Proteomic research, for instance, has revealed dysregulated motor neuron proteins in SMA, providing information on how the disease develops. By identifying biomarkers such as changed amino acid profiles in maple syrup urine disease (MSUD), metabolomics captures the downstream effects of genetic and proteomic changes. These omics layers work together to provide a thorough molecular landscape, but integrating them is crucial to revealing (Johnson, Ivanisevic, & Siuzdak, 2016).

AI Techniques for Integrating Multiple Omics

A variety of multi-omics integration techniques, such as supervised, unsupervised, and deep learning methods, are included in artificial intelligence. Using labeled data, supervised machine learning (ML) algorithms like Random Forest (RF) and Support Vector Machines (SVM) forecast disease outcomes and find biomarkers such as mutated genes in rare cancers (Sharifi-Noghabi, Zolotareva, Collins, & Ester, 2019). Novel disease subtypes can be found by using unsupervised techniques such as principal component analysis (PCA) and clustering, which reveal hidden patterns in unlabeled multi-omics data (Bica, Velickovic, Xiao, & Li, 2018). By extracting latent features, deep learning, in particular, convolutional neural networks (CNNs) and auto encoders perform exceptionally well when handling high-dimensional omics data. For example, Deep Insight improves predictive accuracy by converting tabular omics data into image-like representations for CNN analysis (Shou & West, 2019).

To map gene-protein-metabolite relationships, graph neural networks (GNNs) integrate genomics, proteomics, and metabolomics to model molecular interactions as networks. As demonstrated in studies on cystic fibrosis, transfer learning overcomes data scarcity by applying models developed for common diseases to rare diseases. For rare diseases with small sample sizes, federated learning makes it possible for institutions to collaborate on analysis without disclosing private patient information. Strong predictive modeling and mechanistic insights are made possible by these AI techniques in conjunction with multi-omics data (Wen *et al*., 2023).

Application in the Diagnosis of Rare Diseases

By finding molecular signatures, AI-driven multi-omics integration improves the diagnosis of rare diseases. Machine learning models that combine proteomic (chloride channel dysfunction) and genomic (CFTR mutations) data accurately forecast the severity of cystic fibrosis. DL models that combine transcriptomic and metabolomic profiles find biomarkers for early diagnosis of rare cancers, such as pediatric acute lymphoblastic leukemia (ALL) (Liu & Mei, 2023). AI-analyzed liquid biopsies combine proteomic, metabolomic, and genomic information from blood samples, allowing for the non-invasive diagnosis of uncommon neurological conditions such as Huntington's disease. By capturing cross-layer interactions, these techniques improve diagnostic precision and outperform single-omics methods (Tomczak, Czerwińska, & Wiznerowicz, 2015).

Application in Disease Mechanism Prediction

By simulating molecular networks, AI reveals the mechanisms underlying rare diseases. Therapeutic development in SMA has been guided by the identification of dysregulated pathways, such as SMN protein degradation, by GNNs that integrate transcriptomic and proteomic data (Das *et al*., 2022). AI combines proteomic and metabolomic data to identify metabolite accumulation and enzyme deficiencies in lysosomal storage disorders, such as Gaucher's disease, helping to clarify the course of the illness (Chen *et al*., 2024). Recent studies have used DL to predict the mechanisms of Alzheimer's disease, a rare disease with early onset forms, by combining multi-omics data and identifying pathways related to tau and amyloid-beta. These discoveries show how artificial intelligence (AI) can understand the complex, multi-scale mechanisms underlying rare diseases (Rashid & Selvarajoo, 2024).

Application for the Identification of Therapeutic Targets

For rare diseases with limited treatment options, artificial intelligence (AI) can expedite the drug discovery process. For instance, dysregulated dystrophin pathways in DMD have been found by ML models that integrate proteomic and genomic data, suggesting possible targets for gene therapy. To provide individualized treatments, AI is also utilized in cancer research to evaluate multi-omics data and identify drug resistance mechanisms in uncommon tumors (Lee *et al*., 2024).

Obstacles and Restrictions

Despite its potential, AI-driven multi-omics integration faces several challenges. Data Scarcity: Rare diseases frequently have small patient cohorts, which limit the amount of omics data that can be collected and increase the possibility of overfitting. Heterogeneity: The scale and format differences of multi-omics data from different platforms (e.g., mass spectrometry and WGS) complicate integration. Interpretability: Although DL models are powerful, they are often "black boxes," making it challenging to comprehend the mechanisms (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019). Standardization: Without standardized processes for the creation and examination of omics data, reproducibility is reduced. Computational Resources: Processing high-dimensional multi-omics data requires a significant amount of computational power, which limits accessibility.

Materials and Methods

This study employed a computer system that used artificial intelligence (AI) to aggregate and analyze multi-omics datasets, notably genomes, transcriptomics, proteomics, and metabolomics, to find out what causes rare diseases. We got high-quality datasets that are available to the public from well-known databases such as NCBI GenBank, GEO, PRIDE, and Metabolites. We focused on disorders that are listed in Orphanet and the NIH Rare Disorders Database. The acquired data underwent extensive processing, including normalization, imputation of missing values, and correction of batch effects with techniques such as ComBat. Relevant properties from each omics layer were selected and streamlined for further analysis utilizing techniques such as t-SNE, auto encoders, and principal component analysis (PCA). We used multi-view learning methods, such as similarity network fusion (SNF) and canonical correlation analysis (CCA), to combine the data into a single, unified picture of the biological systems involved. We used supervised machine learning methods, including support vector machines, random forests, and deep neural networks, as well as unsupervised clustering approaches, to find disease-specific patterns and possible subcategories. We also utilized a multimodal deep learning technique to figure out how the multiple omics layers are related to each other. We used the Gene Ontology (GO) and KEGG databases to do pathway enrichment analysis to figure out what the model predictions meant and to highlight the most important biological discoveries. We used explainable AI methods like SHAP and LIME to make the models easier to understand. We checked and rated the performance of each prediction model using metrics including accuracy, F1-score, and AUC-ROC. This study used independent external datasets for validation where it made sense to do so. Because this study solely used publicly available and anonymous data, it didn't need ethical approval.

Ethical Issues

Privacy issues in federated learning and equitable access to AI-driven treatments are unresolved issues. These problems need innovative solutions if AI is to be fully utilized in rare disease research (Vamathevan *et al*., 2019).

Discussion

Research on rare diseases has been revolutionized by AI-driven multi-omics integration, which makes it possible to conduct thorough analyses of proteomics, metabolomics, and genomics. As demonstrated by their high accuracy in predicting disease outcomes in rare cancers and cystic fibrosis, supervised machine learning models like RF and SVM are excellent at finding biomarkers. DL techniques, such as CNNs and autoencoders, uncover new disease subtypes and pathways by identifying intricate patterns in multi-omics data (Wen *et al*., 2023). As shown in SMA and lysosomal storage disorders, GNNs model molecular networks offer insights into gene-protein-metabolite interactions. Data scarcity is addressed by transfer and federated learning, which is crucial for rare diseases with small sample sizes. These developments are consistent with new research showing AI's role in multi-omics for complex diseases that are similar to rare diseases, like cancer and cardiovascular conditions (Chen *et al*., 2024).

But problems still exist. Model generalizability is limited by data scarcity, especially for ultra-rare diseases with fewer than 100 patients globally. As demonstrated by attempts to standardize cancer multi-omics datasets, heterogeneity in omics data necessitates sophisticated preprocessing and normalization techniques (Tomczak *et al*., 2015). Although interpretable latent factors for multi-omics analysis are provided by tools such as SLIDE (Significant Latent Factor Interaction Discovery and Exploration), interpretability is still a challenge. Data aggregation is made easier by standardization initiatives like the Omics Discovery Index (OmicsDI), but wider adoption is necessary (Bertelli *et al*., 2017). Cloud-based solutions are required for computational demands, but strong frameworks are required for ethical considerations like data privacy and fair access.

To improve mechanistic insights, future research should concentrate on creating interpretable AI models, like explainable neural networks. As demonstrated in cancer research, single-cell multi-omics in conjunction with spatial omics may uncover cellular heterogeneity in rare diseases. For rare disease cohorts, collaborative platforms such as federated learning can increase data access while maintaining privacy. Cardiovascular studies have shown that combining multi-omics with clinical data (such as imaging and patient records) may increase the accuracy of diagnosis and prognosis. AI-driven rare disease research is expected to advance more quickly in these directions, opening the door to customized treatments (Yang *et al*., 2024).

Conclusion

By combining genomics, proteomics, and metabolomics, AI-based multi-omics integration provides a revolutionary method for deciphering the mechanisms underlying rare diseases. It fills gaps in the management of rare diseases by facilitating accurate diagnosis, mechanistic insights, and therapeutic target identification. The potential for AI to model intricate molecular interactions is enormous, notwithstanding obstacles like data scarcity, heterogeneity, and interpretability. AI-driven multi-omics has the potential to transform rare disease research by overcoming these obstacles through creative approaches and teamwork, providing individualized treatment options for patients with few other options.

Limitations and Future Research Directions

There are some problems with this study, even though the results are good. First, there aren't enough complete and high-quality multi-omics samples for rare diseases. This could make the AI models less reliable and able to be applied to other situations. Numerous rare illnesses lack sufficient publicly available data, particularly in the realms of proteomics and metabolomics. This may result in data imbalance and an insufficient representation of biology. Second, omics data may be of multiple sizes, dimensions, or be influenced by batch effects, making it difficult to combine. Even with better preprocessing, they may reduce prediction model accuracy and effectiveness. The last thing this study does is give us some ideas about how rare illnesses might work, but it doesn't prove them. More research is needed to back up the computer estimates.

Declarations

Ethical Approval and Consent to Participate: This study strictly adhered to the Declaration of Helsinki and relevant national and institutional ethical guidelines. Informed consent was not required, as secondary data available on websites was obtained for analysis. All procedures performed in this study were by the ethical standards of the Helsinki Declaration.

Consent for Publication: Not Applicable

Availability of Data and Materials: Data for this study will be made available upon request from the corresponding author.

Competing Interest: The authors declare that they have no competing interests.

Funding: Not Applicable

Authors' Contribution: All authors have an active role in the conduct, writing, and submission to the journal.

Acknowledgement: We authors appreciate the assistance of the colleagues, fellows, and respondents of the study for their cooperation in conducting this study.

References

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928), 198-207.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11), 745-755.

Bertelli, C., Laird, M. R., Williams, K. P., Group, S. F. U. R. C., Lau, B. Y., Hoad, G., . . . Brinkman, F. S. (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research*, 45(W1), W30-W35.

Bica, I., Velickovic, P., Xiao, H., & Li, P. (2018). *Multi-omics data integration using cross-modal neural networks*. Paper presented at ESANN.

Boycott, K. M., Rath, A., Chong, J. X., Hartley, T., Alkuraya, F. S., Baynam, G., . . . den Dunnen, J. T. (2017). International cooperation is needed to enable the diagnosis of all rare genetic diseases. *The American Journal of Human Genetics*, 100(5), 695-705.

Chen, Q., Gao, F., Wu, J., Zhang, K., Du, T., Chen, Y., . . . Tang, J. (2024). Comprehensive pan-cancer analysis of mitochondrial outer membrane permeabilisation activity reveals positive immunomodulation and assists in identifying potential therapeutic targets for immunotherapy resistance. *Clinical and Translational Medicine*, 14(6), e1735.

Das, T., Kaur, H., Gour, P., Prasad, K., Lynn, A. M., Prakash, A., & Kumar, V. (2022). Intersection of network medicine and machine learning towards investigating the key biomarkers and pathways underlying amyotrophic lateral sclerosis: a systematic review. *Briefings in Bioinformatics*, 23(6), bbac442.

Haendel, M., Vasilevsky, N., Unni, D., Bologa, C., Harris, N., Rehm, H., . . . McMurry, J. (2020). How many rare diseases are there? *Nature Reviews Drug Discovery*, 19(2), 77-78.

Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, 17(7), 451-459.

Kang, M., Ko, E., & Mersha, T. B. (2022). A roadmap for multi-omics data integration using deep learning. *Briefings in Bioinformatics*, 23(1), bbab454.

Lee, J., Kim, D., Kong, J., Ha, D., Kim, I., Park, M., . . . Kim, S. (2024). Cell-cell communication network-based interpretable machine learning predicts cancer patient response to immune checkpoint inhibitors. *Science Advances*, 10(5), eadj0785.

Liu, X.-Y., & Mei, X.-Y. (2023). Prediction of drug sensitivity based on multi-omics data using deep learning and similarity network fusion approaches. *Frontiers in Bioengineering and Biotechnology*, 11, 1156372.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). *Interpretable machine learning: Definitions, methods, and applications*. arXiv preprint arXiv:1901.04592.

Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., . . . Rath, A. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. European journal of human genetics, 28(2), 165-173.

Rashid, M. M., & Selvarajoo, K. (2024). Advancing drug-response prediction using multi-modal and-omics machine learning integration (MOMLIN): a case study on breast cancer clinical data. *Briefings in Bioinformatics*, 25(4), bbae300.

Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, 49, 107739.

Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., & Ester, M. (2019). MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics, 35(14), i501-i509.

Shou, C., & West, M. (2019). *A Tree-based radial basis function method for noisy parallel surrogate optimization*. arXiv preprint arXiv:1908.07980.

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, 14, 1177932219899051.

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). Review The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, (1), 68-77.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., . . . Spitzer, M. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), 463-477.

van Tuijl, J., van Heck, J. I., Bahrar, H., Broeders, W., Wijma, J., Ten Have, Y. M., . . . Joosten, L. A. (2024). Single high-fat challenge and trained innate immunity: A randomized controlled crossover trial. *Iscience*, 27(11).

Wekesa, J. S., & Kimwele, M. (2023). A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. *Frontiers in Genetics*, 14, 1199087.

Wen, Y., Zheng, L., Leng, D., Dai, C., Lu, J., Zhang, Z., . . . Bo, X. (2023). Deep learning-based multiomics data integration methods for biomedical applications. *Advanced Intelligent Systems*, 5(5), 2200247.

Yang, B., Liu, S., Xie, J., Tang, X., Guan, P., Zhu, Y., . . . Li, W. (2024). Hierarchical learning of gastric cancer molecular subtypes by integrating multi-modal DNA-level omics data and clinical stratification. *Quantitative Biology*, 12(2), 182-196.

Note: **Open Access Public Health and Health Administration Review** is recognized by the Higher Education Commission of Pakistan in the Y category.

Disclaimer/ Publisher's Note: The statements, opinions, and data contained in all publications in this journal are solely those of the individual author(s) and not of the MDPIP and/ or the editor(s). MDPIP and editor(s) disclaim responsibility for any injury to the people or property resulting from any ideas, methods, instructions, or products referred to *in the content.*